Осина Т.Г. ***Украина***

Национальный технический университет «Харьковский политехнический институт»

МОДЕЛИ ДИСКРИМИНАЦИИ В УСЛОВИЯХ НЕОПРЕДЕЛЕННОСТИ КОНТРОЛИРУЕМЫХ ПАРАМЕТРОВ

Пусть N – общее количество нестандартных образцов, предназначенных для градуировки (обучения) системы измерительного контроля уровней параметра Y. Если p – число показателей контроля (сигналов X_1, …X_p, несущих информацию о значениях уровней параметра Y), то для двух ($r = \overline{1, 2}$) уровней дискриминации ожидаемое количество измерительной информации определяется выражением

$$I = \log\sqrt{1+\left(\frac{\sigma_X}{\sigma_{\Delta X}}\right)^2},$$

где σ_x^2 – дисперсия функции $g_2(x)$ до изменения; $\sigma_{\Delta x}^2$ – дисперсия функции $g_2(x)$ после изменения.

Количество информации I об уровнях параметра Y тем больше, чем больше отношение $\left(\sigma_x / \sigma_{\Delta x}\right)^2$.

Дисперсия σ_x^2 полностью определяется векторами средних $\overline{X}^{(r)}$, $r = \overline{1, 2}$ и диагональной матрицей D и в скалярном виде может быть представлена как дисперсия дискретной случайной величины $\xi_r = M\left[g(x)/y_r\right]$, имеющей нулевое среднее и одинаковые, по 0,5, вероятности появления уровней y_1 и y_2 параметра контроля Y:

$$\sigma_X^2 = 0.5\cdot\left\{M\left[g_2(x)/y_1\right]\right\}^2 + 0.5\cdot\left\{M\left[g_2(x)/y_2\right]\right\}. \quad (1)$$

Так как

$$\left\{M\left[g_2(x)/y_1\right]\right\} = \left\{M\left[g_2(x)/y_2\right]\right\} = 0{,}5\sum_{i=1}^{p}\tilde{\delta}_i^2, \quad (2)$$

то при априорно заданных средних $M_i^{(r)}$ и дисперсиях σ_i^2 показателей $\{X_i\}$ выражение (1) запишется в виде:

$$\sigma_x^2 = \frac{1}{4}\left[\sum_{i=1}^{p}\left(\frac{\mu_i^{(1)}-\mu_i^{(2)}}{\sigma_i}\right)^2\right]^2. \quad (3)$$

Дисперсия $\sigma_{\Delta x}^2$ – это центральный момент второго порядка функции $g_2(x)$, когда в качестве элементов векторов средних $X^{(r)}$ и диагональных матриц D_r используются их оценки $\overline{x}_i^{(r)}$ и $\overline{D}_i^{(r)}$, $i = \overline{1, n_r}$, где n_r – объем обучающей выборки по классу $\pi_r\left(Y \in y_r\right)$:

$$\sigma_{\Delta X}^2 = \sum_{i=1}^{p}\frac{\left(\overline{x}_i^{(1)} - \overline{x}_i^{(2)}\right)^2}{D_i}. \quad (4)$$

Дисперсия $\sigma_{\Delta x}^2$ является случайной величиной $\xi = \sum_{i=1}^{p}\xi_i$ в силу случайного характера оценок $\overline{x}_i^{(1)}$, $\overline{x}_i^{(2)}$ и D_i.

Если объемы обучающих выборок по уровням y_1 и y_2 параметра контроля Y одинаковы $\left(n_r = n, r = \overline{1, 2}\right)$, то каждую из случайных величин ξ_i можно рассматривать как линейно преобразованную случайную величину, имеющую нецентральное F–распределение с одной и (n-1) степенями свободы и параметром нецентральности

$$\lambda_i = \varepsilon_i^2\left(n/2\right), \quad (5)$$

$$\varepsilon_i^2 = \left(\mu_i^{(1)} - \mu_i^{(2)}\right)^2 / D_i \quad (6)$$

Коэффициент линейного преобразования равен (2/n), а величина ξ_i соответствует вероятностной модели $\xi_i \square \left(2/n\right)\cdot F_{1,(n-1),\lambda_i}$.

Математическое ожидание величины ξ_i равно:

$$\chi_{1i} = \left(\frac{n-1}{n-3}\right)\left(\frac{2}{n}+\varepsilon_i^2\right), \quad (7)$$

а дисперсия $\sigma_{\Delta x}^2$, отражающая остаточную неопределенность (энтропию) после измерения уровня параметра Y запишется как

$$\sigma_{\Delta x}^2 = \sum_{i=1}^{p}\tau_{1i}. \quad (8)$$

С учетом (6) и (7) количество ожидаемой измерительной информации, полученной в результате принятия одного из двух решений, равно

$$I = \frac{1}{2}\log\left\{1 + \frac{(n-3)\sum_{i=1}^{p}\varepsilon_i^2}{(n-1)\left[4 + \frac{8p}{n\sum_{i=1}^{p}\varepsilon_i^2}\right]}\right\}. \qquad (9)$$

Из выражения (9) видно, что количество информации I растет при увеличении объема обучающей выборки n и увеличении суммы

$$\varepsilon_p^2 = \sum_{i=1}^{p}\varepsilon_i^2 ,$$

характеризующей общую дискриминирующую способность системы из P информационных показателей контроля.

С другой стороны, если с ростом P числа этих показателей ε_p^2 остается неизменной, количество информации падает.

Таким образом, выбор числа показателей контроля является оптимизационной задачей. Это хорошо иллюстрируется таблицей 1 и графиками, представленными на рисунке 1, которые показывают зависимость количества информации от числа показателей контроля при разных объемах выборок n, когда дискриминирующий свойства показателей X_1, …X_p разные (величина ε_i^2 показателя X_i определяется как $\varepsilon_i^2 = \varepsilon_{i-1}^2 / 2$). Это соответствует существенной информационной неоднородности показателей.

Таблица 1

Количество информации (в бит) для системы из P показателей при разных n ($\varepsilon_1^2 = 1; \varepsilon_2^2 = 0.5, \varepsilon_3^2 = 0.25...$)

n	P							
	1	2	3	4	5	6	7	опт. P
60	0,1563	0,2213	0,2498	0,2618	0,2590	0,2559	0,2527	P=4
500	0,1604	0,2287	0,2603	0,2754	0,2756	0,2754	0,2751	P=5

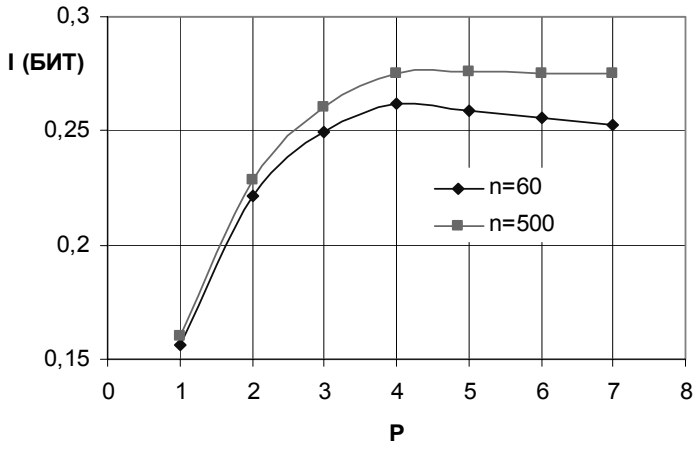

Рис. 1 Количество ожидаемой измерительной информации в функции P и n.

Из таблицы 1 видно, что максимальное количество информации при n=60 соответствует числу признаков P=4 (I=0,2618), а максимальное количество информации при n=500 соответствует P=5 (I=0,2756). Из таблицы 1 следует, также, что максимум информации растет как при увеличении объема выборки n, так и числа показателей контроля P.

Из анализа выражения (9) также следует, что количество информации снижается при увеличении отношения P/n, что указывает на необходимость повышения n при увеличении размерности системы информационных показателей контроля (условия выборки n_{min}– сохранение хотя бы постоянства отношения P/n).